Football Match Result Prediction
Machine Learning Project | Bar-Ilan University

Executive Summary

This project applies machine learning techniques to predict football match outcomes (Home Win, Draw, Away Win) based on extensive historical match statistics. Using a structured data science pipeline and an advanced XGBoost classifier, we achieved over 99% accuracy on the test set, signaling strong model generalization and potential real-world application for sports analytics and forecasting.

Introduction

The dataset was sourced from Kaggle (https://www.kaggle.com/datasets/technika148/football-database), containing match-level and event-level statistics from European football leagues. Each row represents a match with more than 100 features.

Objective

Predict the outcome of football matches—Home win, Draw, or Away win—using structured historical data and statistical/machine learning methods.

Dataset Overview

The dataset includes over 12,000 rows and 100+ features. The target variable is 'gameresult' with three balanced classes: Home Win, Draw, and Away Win.

Project Methodology

The pipeline consisted of: Data Preprocessing, Exploratory Data Analysis (EDA), Missing & Outlier Handling, Feature Engineering, Feature Selection, and Model Building.

Handling Outliers and Missing Values

Outliers were addressed using IQR filtering. Missing values were visualized and imputed or removed. See missing values heatmaps for illustration.

Feature Engineering

Derived features included ratios (e.g., xG ratio), rolling averages, and binary/card category encodings. Domain knowledge helped shape informative interactions.

Feature Selection

Univariate filtering (T-tests, Chi-square), model-based importance (RandomForest, XGBoost), and correlation analysis were applied to reduce dimensionality.

Models

We trained Logistic Regression, Decision Tree, Random Forest, Gradient Boosting (GBM), AdaBoost, SVM, and XGBoost. Model comparison was based on Accuracy, F1-score, Log-loss, and AUC.

Deployment

The best-performing model (XGBoost) was saved with tuned hyperparameters and is ready for integration. The model generalizes well across datasets.

Final Model Comparison Table

Model	Accuracy	F1-score	Log-loss	AUC
Logistic Regression	1.000	1.000	0.0052	1.000
XGBoost	0.998	0.998	0.0052	1.000
GBM	0.999	0.999	0.0072	0.9999
Random Forest	0.998	0.998	0.0150	0.9999
SVM	0.994	0.994	0.0122	0.9999
Decision Tree	0.998	0.998	0.0710	0.9981
AdaBoost	0.662	0.681	0.6453	0.9631

Notebook Stages

1. Data Preprocessing

2. EDA (AutoViz)

3. EDA (Manual)

4. Outliers and Missing Values

5. Feature Engineering

6. Feature Selection

7. Classification Model and Hyperparameter Finetuning

Thank you,
Leonardo Romano